Safe and effective importance sampling

Art Owen
Stanford University

Yi Zhou
Goldman-Sachs

This version: March 1999

Abstract

We present two improvements on the technique of importance sampling. First we show that importance sampling from a mixture of densities, using those densities as control variates, results in a useful upper bound on the asymptotic variance. That bound is a small multiple of the asymptotic variance of importance sampling from the best single component density. This allows one to benefit from the great variance reductions obtainable by importance sampling, while protecting against the equally great variance increases that might take the practitioner by surprise. The second improvement is to show how importance sampling from two or more densities can be used to approach a sampling variance of zero even for integrands that take both positive and negative values.

KEY WORDS: control variates, Monte Carlo, rare events, reliability, value at risk, variance reduction
1 Introduction

We consider here the problem of computing

\[ I = \int_{\mathcal{D}} f(x)q(x)dx, \quad (1) \]

where \( \mathcal{D} \subseteq \mathbb{R}^d \) is the domain of \( x \) and \( q(x) \) is a probability density function on \( \mathcal{D} \). We are especially interested in problems where \( f \) is a spiky function, by which we mean that an appreciable fraction of the variance of \( f \) may be attributed to a subset of \( \mathcal{D} \) having relatively small probability under sampling from \( q \). Spiky integrands of this sort arise in high energy physics, Bayesian statistics, computational finance, computer graphics, and in the computation of rare event probabilities, as for example in reliability problems.

It is natural to consider Monte Carlo sampling for these problems, especially when \( d \) is not small. Furthermore, for spiky integrands it is common to employ some form of importance sampling, in order to get more data from the important region containing the spikes.

This paper considers two improvements in importance sampling. The first is importance sampling from a mixture density, while employing the mixture components as control variates. We show that this method is, asymptotically, not much worse than importance sampling from the best of the mixture components, even if all but one of those components would have given an infinite variance. The practical benefit is that one can employ several densities in the hopes that one of them is particularly good, without losing too much if one or more of them is particularly bad. This method is also useful when several integrals, to be estimated from the same sample, have different important regions.

For the second improvement, we introduce a technique of positivisation. The integrand is split into nonnegative and nonpositive parts and importance sampling is applied to each of them. This allows one to approach zero
variance even for integrands taking both signs. One can also positivise the difference \( f - h \) between the integrand and a control variate. The method combines naturally with the mixture sampling described above.

For definiteness, our discussion is couched in terms of \( X \) having density \( q \) over \( D \subseteq \mathbb{R}^d \). Our main results extend to discrete random variables \( X \), upon replacing densities by probability mass functions.

### 1.1 Outline

In Section 2 we review importance sampling, and show how it can succeed spectacularly when \( p \) is nearly proportional to \( qf \). But in Example 1 given there, importance sampling fails spectacularly even though \( p \) is nearly proportional to \( qf \) in the important region. The cause is lack of proportionality \textit{outside} of the important region. The method of defensive importance sampling (Hesterberg 1995) addresses this problem, by sampling from a mixture of \( p \) and \( q \). But defensive importance sampling can cause a great deterioration when ordinary importance sampling would have worked. See Example 2 there.

It is possible to get the best of both worlds, benefiting from importance sampling when it works well, while protecting against its failings. This safe and effective importance sampling can be obtained by employing \( p \) and \( q \) as control variates while sampling from a mixture of \( p \) and \( q \). Section 3 presents the method of control variates as it is used in combination with importance sampling. Theorem 1 there gives conditions under which estimated control variate coefficients are essentially equivalent to the optimal ones.

Section 4 presents a method of importance sampling from a finite mixture of densities. There we show that using the individual mixture components as control variates yields an asymptotic variance not much, if any, larger than that of importance sampling from single best mixture component. Theorem 2
proves this assuming optimal control variate coefficients, and Section 4.1 gives mild conditions under which we can expect our sample coefficients to behave like the optimal ones.

Section 5 presents the method of multiple importance sampling due to Veach & Guibas (1995). Like defensive importance sampling, multiple importance sampling is motivated by the desire to pool importance sampling methods and get nearly the best performance.

Section 6 is a simulation of Examples 1 and 2. The proposed hybrid methods are nearly best on both examples.

Section 7 introduces a version of multiple importance sampling that can approach zero variance on integrands taking both signs. We show there how to exploit a control variate in conjunction with multiple importance sampling.

Section 8 gives an example in which positivisation and multiple importance sampling are combined. In this example, we replace the integrand by a spiky one and obtain a large variance reduction by mixture sampling.

1.2 Background

An early use of mixture sampling is Torrie & Valleau (1977). Arsham, Fuerverger, McLeish, Kreimer & Rubinstein (1989) advocate the use of the likelihood ratio as a control variate in importance sampling. Hesterberg (1995) considers the use of this ratio as a control variate in combination with defensive mixture sampling, and obtains an asymptotic bound on the ratio of the resulting variance to that of sampling from the nominal density. This appears to be the first published result of this kind, though it appeared earlier in a dissertation (Hesterberg 1988).

Our contribution is to extend the theoretical result to more general mixture samples, to describe conditions under which estimated coefficients approach the true ones, and to combine it with positivisation.
Multiple importance sampling was proposed by Veach & Guibas (1995) for problems in computer graphics. Our positivisation technique is based upon it.

2 Importance sampling

This section reviews importance sampling and introduces our notation. Integrals without an explicit domain are assumed to be over $\mathcal{D}$. For brevity, we sometimes omit the arguments of functions, for example in writing $q$ for $q(x)$, when the argument $x$ is clear from context.

2.1 Basic importance sampling

In importance sampling, we sample $X_i$ independently from a density $p$, with $p(x) > 0$ whenever $f(x)q(x) \neq 0$, then estimate $I$ by

$$\hat{I}_p = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)q(X_i)}{p(X_i)}.$$  \hspace{1cm} (2)

It is easy to see that $E(\hat{I}_p) = I$. There are two practical constraints on importance sampling: it must be feasible to sample from $p$, and we must be able to compute $fq/p$. For the second constraint, assuming we can compute $f$, it suffices to be able to compute the ratio $q/p$, which can be simpler than computing $p$ and $q$ separately.

By elementary manipulations we find $\text{Var}(\hat{I}_p) = \sigma_p^2/n$, where

$$\sigma_p^2 = \int \left( \frac{f(x)q(x)}{p(x)} - 1 \right)^2 p(x)dx = \int \frac{f^2(x)q^2(x)}{p(x)}dx - I^2$$  \hspace{1cm} (3)

is referred to as the asymptotic variance. The density $p^*$ that minimizes this asymptotic variance is known to be proportional to $|f(x)|q(x)$ (Kahn &
In the common special case with $f(x) \geq 0$ and $I > 0$, we may write $p^*(x) = f(x)q(x)/I$. This density gives $\sigma_p^2 = 0$, but does not satisfy the practical constraint, because $fq/p^* = I$, and $I$ is unknown.

The practical value of knowing that $p^* = fq/I$ is optimal, is that it suggests that a good importance sampling density should be roughly proportional to $fq$. Let $r(x) = f(x)q(x) - Ip(x)$. This $r(x)$ is the residual from proportionality, and it satisfies $\int r(x)dx = 0$. We can re-write (3) as

$$\sigma_p^2 = \int \frac{r(x)^2}{p(x)}dx. \quad (4)$$

### 2.2 Failure of importance sampling

Importance sampling can fail dramatically, even when $p$ matches $fq/I$ well near the mode(s) of $fq/I$. The cause is the appearance of $p(x)$ in the denominator of (3) and (4). If $p(x)$ decreases towards zero faster than $f^2(x)q^2(x)$ as $x$ moves away from its mode(s), we can find that $\text{Var}(\hat{I}_p) = \infty$. The irony is that this infinite variance may be due to a region of $\mathcal{D}$ that is unimportant in ordinary Monte Carlo sampling.

To illustrate this point, we present Example 1 below. We use the beta density function

$$B(x, a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)/\Gamma(a + b)}, \quad 0 < x < 1$$

where $a > 0$ and $b > 0$ are parameters and $\Gamma(z)$ is the gamma function. For vectors in the unit cube, we use superscripts to denote their components.

**Example 1** Let $\mathcal{D} = (0, 1)^5$, with nominal density $q(x) = U(0, 1)^5$. The
Figure 1: Shown are plots of the integrand $f_1(x)$ (solid) and the density $p(x)$ (dashed) along two transects through the unit cube $(0,1)^5$. In the left panel $x = (z,0.5,0.5,0.5,0.5)$, while in the right panel $x = (z,z,z,z,z)$, both for $0 < z < 1$. The density is nearly proportional to the integrand near the mode, but gives an infinite variance if used in importance sampling.

The integrand is

$$f_1(x) = 0.9 \times \prod_{j=1}^{5} B(x^j, 20, 20) + 0.1 \times \prod_{j=1}^{5} B(x^j, 2, 2), \quad (5)$$

and the importance sampling density is

$$p(x) = \prod_{j=1}^{5} B(x^j, 20, 20). \quad (6)$$

It is easy to show that $I = 1$ and $\sigma_p^2 = \infty$ for Example 1. As Figure 1 shows, this density is very nearly proportional to $f_1q$, so we might have expected a good result. In similar plots using $f_1$ and $0.9p$, the curves are visually indistinguishable.

In Example 1 importance sampling fails despite having $p$ nearly propor-
tional to \(f_1q\) in the important part of the domain. We return to this example in the simulations of Section 6.

2.3 Defensive importance sampling

The failure of importance sampling can be countered by defensive importance sampling (Hesterberg 1995). Let \(p(x)\) be a density that is thought to be a good approximation to \(fq/I\), at least in the important part of \(D\). Pick \(\alpha_1\) with \(0 < \alpha_1 < 1\) and define the mixture density

\[
p_\alpha(x) \equiv \alpha_1 q(x) + \alpha_2 p(x)
\]

(7)

where \(\alpha_2 = 1 - \alpha_1\). Here \(\alpha\) is the vector \((\alpha_1, \alpha_2)\).

By mixing in some of \(q\) we prevent \(p_\alpha\) from being much smaller than \(q\) anywhere. We find that

\[
\sigma^2_{p_\alpha} = \int \frac{f^2(x)q^2(x)}{p_\alpha(x)} dx - I^2
\]

\[
\leq \frac{1}{\alpha_1} \left( \int f^2(x)q(x) dx \right) - I^2
\]

\[
= \frac{1}{\alpha_1} \left( \sigma^2_q + I^2 \alpha_2 \right).
\]

(8)

Equation (8) provides a kind of insurance against the worst effects of importance sampling. If the nominal density provides a finite variance, then defensive importance sampling will too.

Hesterberg (1995) recommends using \(\alpha_1\) between 0.1 and 0.5. Spiky non-negative integrands can have \(\sigma_q \gg I > 0\), in which case this advice will, approximately, bound the sampling variance by between 2 and 10 times what it would be under the nominal density.
2.4 Failure of defensive importance sampling

Defensive importance sampling can greatly increase the variance over what it would have been with ordinary importance sampling. In other words, the premium we pay for the insurance (8) can be very high. The root of the problem is that if $p$ is nearly proportional to $fq$ then $\alpha_1q + \alpha_2p$ will not be, apart from trivial cases with $q$ nearly proportional to $fq$.

Here is an example where defensive importance sampling fails, in that it destroys the near proportionality of the original non-defensive method.

**Example 2** Let $D = (0,1)^5$, with nominal density $q(x) = U(0,1)^5$. The integrand is

$$f_2(x) = \prod_{j=1}^{5} B(x^j, 20, 20) + 0.1 \prod_{j=1}^{5} \sin \left( 60\pi(x^j - 1/3) \right) 1_{1/3 \leq x^j \leq 2/3}. \tag{9}$$

and the importance sampling density is (6), from Example 1.

The integral here is $I = 1$ and the functions $f_2$ and $p$ are visually indistinguishable. Example 2 is considered further in Section 6, where defensive importance sampling, as expected, greatly increases the variance.

In some settings we can show that as $p$ differs from proportionality by $O(\epsilon)$ that defensive sampling causes a loss of efficiency of $O(\epsilon^{-2})$ as $\epsilon \to 0$. The next example provides a simple illustration of this effect.

**Example 3** Let $D = (0,1)$, with nominal density $q = U(0,1)$. The integrand is $f(x) = x^a$, for some $a > 1$ and $p(x) = (b+1)x^b$, for $b > 1$.

In Example 3 the integrand and the density both have a spike at 1. Easy calculations show that $\sigma_p^2 = O((b-a)^2)$ as $b \to a$, so that if one guesses a very good value for $b$ then a very tiny variance results. But for $\alpha_1 > 0$, we find $\sigma_p^2$ tends to a nonzero limit as $b \to a$. The result is a variance that is
larger by $O((b - a)^{-2})$ than what we could have achieved without defensive sampling.

3 Control variates with importance sampling

The method of control variates uses knowledge of one or more integrals to reduce variance in the estimate of $I$. The basic method is described in texts such as Ripley (1987), Bratley, Fox & Schrage (1987), and Hammersley & Handscomb (1964). Here we present control variates in combination with importance sampling.

Suppose that we know $\int h_j(x)dx = \mu_j$, $j = 1, \ldots, m$. We assume that $p(x) > 0$ if any $h_j(x) > 0$, or if $f(x)q(x) > 0$. Let $\beta = (\beta_1, \ldots, \beta_m)$ be a vector of real values. Under independent sampling of $X_i$ from $p(x)$,

$$
\hat{I}_{p, \beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)q(X_i) - \sum_{j=1}^{m} \beta_j h_j(X_i)}{p(X_i)} + \sum_{j=1}^{m} \beta_j \mu_j
$$

(10)

is an unbiased estimate of $I$.

The variance of $\hat{I}_{p, \beta}$ is $\sigma^2_{p, \beta}/n$, where

$$
\sigma^2_{p, \beta} = \int \left( \frac{f(x)q(x) - \sum_{j=1}^{m} \beta_j h_j(x)}{p(x)} - I + \sum_{j} \beta_j \mu_j \right)^2 p(x)dx.
$$

(11)

Let $\beta^*$ minimize the integral in (11), over $\beta$, for the given functions $f$, $p$, and $q$. Equation (11) suggests that an estimate $\hat{\beta}$ of $\beta^*$ can be found by a multiple regression (including an intercept term) of $f(X_i)q(X_i)/p(X_i)$ on predictors $h_j(X_i)/p(X_i)$.

Because the integral includes an intercept coefficient $\hat{\beta}_0$, the residuals will
sum to zero. As a result equation (10) with $\beta = \hat{\beta}$ simplifies to

$$\hat{I}_{p,\beta} = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j \mu_j.$$  

**Theorem 1** Suppose that there is a unique vector $\beta^*$ that minimizes $\sigma_{p,\beta}^2$, and let $\hat{\beta}$ be determined by least squares as described above. Suppose further that the expectations under sampling from $p$ of the following quantities exist and are finite: $h_j^2 h_l^2 / p^4$, and $h_j^2 f^2 q^2 / p^4$, for $1 \leq j \leq l \leq m$. Then

$$\hat{\beta}_j = \beta^*_j + O_p(n^{-1/2}),$$  

for $j = 1, \ldots, m$, and

$$\hat{I}_{p,\beta} = \hat{I}_{p,\beta^*} + O_p(n^{-1}).$$  

**Proof:** Each $\hat{\beta}_j$ is a smooth function of means, in this case the crossproducts employed in the regression. The assumed moments ensure that sample versions of these means converge at the $O(n^{-1/2})$ rate. The uniqueness of $\beta^*$ rules out singularity of the regression, so that the $O(n^{-1/2})$ rate holds also for $\hat{\beta}_j$, and thus (12) holds.

To establish equation (13), write $\hat{I}_{p,\beta} - \hat{I}_{p,\beta^*} = \sum_{j=1}^{m} (\hat{\beta}_j - \beta^*_j)(\mu_j - \hat{\mu}_j)$ where $\hat{\mu}_j = n^{-1} \sum_{i=1}^{n} h_j(X_i) / p(X_i) = \mu_j + O_p(n^{-1/2}). \square$

Theorem 1 shows that while $\hat{\beta}_j$ approaches $\beta^*_j$ at the standard Monte Carlo rate, the effect of substituting $\hat{\beta}_j$ for unknown optimal $\beta^*_j$ is asymptotically negligible. For this reason it is customary to analyze control variate methods as if the unknown optimal values were being used, while in practice one uses the estimated coefficients. This practice is usually reasonable in Monte Carlo sampling where $n$ is large. Because $\hat{I}_{p,\beta} - \hat{I}_{p,\beta^*}$ is a sum of $m$ terms, $n$ should be large compared to $m$, as it typically is.
4 Mixture sampling

Suppose we have a list of density functions \( p_j, j = 1, \ldots, m \). In defensive importance sampling the \( p_j \) include the nominal density and another thought to be nearly proportional to \( f q \). In other settings we may have a list of densities of which we hope that one or more is roughly proportional to \( f q \). Finally, we may have more than one integrand to consider in our simulation, and each \( p_j \) may be customized for a subset of these integrands. These densities may have been suggested by subject matter knowledge, or they may have been found by numerical search.

We will sample from the mixture density \( p_\alpha(x) = \sum_{j=1}^{m} \alpha_j p_j(x) \) where \( \alpha_j > 0 \) and \( \sum_{j=1}^{m} \alpha_j = 1 \). Because \( f p_j(x) dx = 1 \), and \( p_\alpha(x) > 0 \) whenever \( p_j(x) > 0 \), we can use the \( p_j \) as control variates as described in Section 3. We write the resulting estimator as

\[
\tilde{I}_{\alpha, \beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)q(X_i) - \sum_{j=1}^{m} \beta_j p_j(X_i)}{\sum_{j=1}^{m} \alpha_j p_j(X_i)} + \sum_{j=1}^{m} \beta_j,
\]

(reserving the notation \( \tilde{I}_{\alpha, \beta} \) for deterministic mixture sampling, introduced in Section 4.3).

The asymptotic variance \( \sigma_{p_\alpha, \beta}^2 \) of \( \tilde{I}_{\alpha, \beta} \) is given by (11) with \( h_j = p_j \) and \( \mu_j = 1 \). This is a positive semi-definite quadratic function in \( \beta \). The minimum is not unique, as required by Theorem 1. For any scalar \( c \), we have \( \tilde{I}_{p_\alpha, \beta+c\alpha} = \tilde{I}_{p_\alpha, \beta} \), and so if \( \beta^* \) is a minimizer of \( \sigma_{p_\alpha, \beta}^2 \), then so is \( \beta^* + c\alpha \). Section 4.1 describes how to apply Theorem 1 to this setting.

Theorem 2 below shows that \( \tilde{I}_{\alpha, \beta} \) is unbiased, and that for an optimal \( \beta \), the variance is never larger than what one gets from an importance sample of size \( n\alpha_j \) from \( p_j \).

**Theorem 2** Let \( p_j, \alpha_j \) and \( \tilde{I}_{\alpha, \beta} \) be as above. If at least one of \( p_j(x) > 0 \) whenever \( f(x)q(x) > 0 \), then for any \( \beta \), we have \( E(\tilde{I}_{\alpha, \beta}) = I \).
Let $\sigma_{p_{\alpha},\beta}^2$ be the asymptotic variance of $\tilde{I}_{\alpha,\beta}$ and let $\sigma_{p_j}^2$ be the asymptotic variance ($\beta$), under importance sampling from $p_j$. If $\beta^*$ is any minimizer of $\sigma_{p_{\alpha},\beta}^2$, then

$$
\sigma_{p_{\alpha},\beta}^2 \leq \min_{1 \leq j \leq m} \alpha_j^{-1} \sigma_{p_j}^2; 
$$

(15)

**Proof:** To establish unbiasedness, write

$$
E(\tilde{I}_{\alpha,\beta}) = \int \frac{f(x)q(x) - \sum_{j=1}^{m} \beta_j p_j(x)}{p_\alpha(x)} p_\alpha(x) dx + \sum_{j=1}^{m} \beta_j = I.
$$

Next we prove that $\sigma_{p_{\alpha},\beta}^2 \leq \sigma_{p_1}^2/\alpha_1$. Consider the vector $\beta$ with $\beta_1 = 0$, and $\beta_j = -I\alpha_j/\alpha_1$ for $j > 1$. Let $r_1(x) = f(x)q(x) - Ip_1(x)$, so $\int r_1(x) dx = 0$. Substituting these values, we find

$$
\sigma_{p_{\alpha},\beta}^2 = \int \left( \frac{f(x)q(x) + I\alpha_1^{-1} \sum_{j=2}^{m} \alpha_j p_j(x)}{p_\alpha(x)} - I - I\alpha_1^{-1} \sum_{j=2}^{m} \alpha_j \right)^2 p_\alpha(x) dx
$$

$$
= \int \left( \frac{Ip_1(x) + r_1(x) + I\alpha_1^{-1}(p_\alpha(x) - \alpha_1 p_1(x))}{p_\alpha(x)} - I\alpha_1^{-1} \right)^2 p_\alpha(x) dx
$$

$$
= \int \frac{r_1(x)^2}{p_\alpha(x)} dx
$$

$$
\leq \frac{1}{\alpha_1} \int \frac{r_1(x)^2}{p_1(x)} dx
$$

$$
= \frac{\sigma_{p_1}^2}{\alpha_1},
$$

using formula (4). Now $\sigma_{p_{\alpha},\beta}^2 \leq \sigma_{p_{\alpha},\beta}^2 \leq \alpha_1^{-1} \sigma_{p_1}^2$. By making similar arguments for $j = 2, \ldots, m$, equation (15) is established. □

### 4.1 Estimating the coefficients

We now turn to the problem of finding good control variate coefficients $\beta_j$ to use in $\tilde{I}_{p_{\alpha},\beta}$. Theorem 1 provides sufficient conditions under which esti-
mated control variate coefficients are essentially as good as the optimal ones: uniqueness of the minimizer, bounded expectations for \( p_j^2 p_k^2 / p_\alpha^4 \), and bounded expectations for \( p_j^2 f^2 q^2 / p_\alpha^4 \).

Because \( \sum_j \alpha_j p_j(x) / p_\alpha(x) = 1 \) for all \( x \), the regression described in Section 3 is singular, and special care must be taken. Let us suppose affine independence of the \( p_j \), that is if \( \gamma_0 + \sum_{j=1}^m \gamma_j p_j(x) = 0 \) for all \( x \), then all \( \gamma_j = 0 \). Under this condition dropping one control variates \( p_j / p_\alpha \), from the regression is equivalent to selecting the unique minimizer of (11) over \( \beta_j \), \( j \neq j' \), with \( \beta_j = 0 \). We could also drop the intercept term.

Because \( p_j^2(x) p_k^2(x) / p_\alpha^4(x) \leq \alpha_j^{-2} \alpha_k^{-2} \) for all \( x \), the first moment bound follows easily. Now suppose that for at least one of the \( p_j \), the asymptotic variance \( \sigma_{p_j}^2 \) from (3) is finite. Then
\[
\int \frac{p_j^2 f^2 q^2}{p_\alpha^4} dx \leq \alpha_k^{-2} \alpha_j^{-1} \int \frac{f^2 q^2}{p_j} dx = \alpha_k^{-2} \alpha_j^{-1} (I^2 + \sigma_{p_j}^2) < \infty.
\]

Thus, if the densities \( p_j \) are affinely independent and at least one of them gives rise to a finite importance sampling variance, we can expect estimated control variate coefficients to behave like the optimal ones in the way described by Theorem 1.

We prefer to use a singular value decomposition (SVD) to compute the regression coefficients \( \hat{\beta}_j \). See Golub & Van Loan (1983). The \( \beta \) vector computed by the SVD can be expected to differ by some multiple of \( \alpha \) from the one obtained by dropping one or more predictors, but this does not affect the estimate \( \hat{I}_{p_\alpha, \beta} \). The SVD effectively drops a linear combination of predictors from the regression.

The SVD will still work if there are additional dependencies among the regressors \( p_j / p_\alpha \), that would have required dropping two or more predictors. Such additional singularities might arise by accident or design, as for example if one of the \( p_j \) is a mixture of some others.
4.2 Interpretation of Theorem 2

Theorem 2 shows that, with an optimal coefficient vector $\beta$, we get an asymptotic variance that is at least as good as we would have had with $n\alpha_j$ observations from $p_j$. It is hard to expect that we could do better in general, because we get on average $n\alpha_j$ observations from $p_j$, and $p_j$ may be the only one of our component densities that would have given a finite asymptotic variance in importance sampling.

For defensive importance sampling, we find that $\sigma_{a,\beta}^2 \leq \sigma_q^2 / \alpha_1$ as insurance, removing the influence of $I$ in (8). The premium we pay is bounded because $\sigma_{a,\beta}^2 \leq \sigma_p^2 / \alpha_2$. We can also find this by reversing the roles of $p$ and $q$ in defensive importance sampling with a control variate.

There is still the worry that one or more bad sampling densities $p_j$ could distort the sample value $\hat{\beta}$ enough to destroy the asymptotic equivalence of $\tilde{I}_{a,\beta}$ and $\tilde{I}_{a,\beta'}$. The discussion in Section 4.1 shows that this will not happen as long as at least one of the component densities has $\sigma_{p_j}^2 < \infty$.

4.3 Deterministic mixture sampling

In a deterministic mixture sample, one takes $n_j = n\alpha_j$ observations (or an integer close to $n\alpha_j$) from the density $p_j$. Let $X_{ji} \sim p_j$ be independent, for $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$. Incorporating control variates $p_j$ the resulting estimate is,

$$
\hat{I}_{a,\beta} = \frac{1}{n} \left( \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{f(X_{ji}) q(X_{ji}) - \sum_{k=1}^{m} \beta_k p_k(X_{ji})}{p_a(X_{ji})} \right) + \sum_{j=1}^{m} \beta_j, \tag{16}
$$

where $X_{ji}$ are independent and drawn from $p_j(x)$.

The estimate (16) is unbiased. We prefer deterministic mixture sampling to ordinary mixture sampling, because, as Hesterberg (1995) shows, it has smaller variance. In our setting, for any $\beta$, we have $\text{Var}(\hat{I}_{a,\beta}) \leq \text{Var}(\tilde{I}_{a,\beta})$. 

15
The unbiasedness of estimates from deterministic mixture sampling extends to the sample moments and cross-moments of the control variates $p_k/p_\alpha$ and of $f_q/p_\alpha$. It follows that the regression procedure for estimating $\beta$, will, under deterministic mixture sampling, estimate a $\beta$ optimal for random mixture sampling. It will even estimate it more effectively than would random mixture sampling. While deterministic mixture sampling improves upon random mixture sampling, there remains the possibility of further improvement by devising an estimate of the possibly different vector $\beta$ that would be optimal for deterministic sampling.

5 Multiple Importance Sampling

Multiple importance sampling is introduced in Veach & Guibas (1995). They consider path sampling algorithms for rendering images in computer graphics. One of their goals is to combine several importance sampling strategies in order to get performance nearly equal to whichever of the strategies is best. In their examples, the optimal method for rendering an image can differ from pixel to pixel in an unpredictable way.

Let $w_j(x)$, $j = 1, \ldots, m$ be a partition of unity: for every $x \in (0,1)^d$, $0 \leq w_j(x) \leq \sum_{j=1}^{m} w_j(x) = 1$. Define

$$\hat{I}_{n,w} = \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} w_j(X_{ji}) \frac{f(X_{ji})}{p_j(X_{ji})},$$

where $X_{ji}$ are independent draws from $p_j$, and the subscripts on $I$ denote the partition of unity and the sample sizes used. The estimate $\hat{I}_{n,w}$ is unbiased under mild conditions on the supports of the function $p_j$ and $w_j$.

Veach & Guibas (1995) consider several ways of selecting $w_j$. Their motivation is to make $fw_j$ “locally proportional” to $p_j$. Their balance heuristic
takes weights
\[ w_j(x) = \frac{n_j p_j(x)}{\sum_{k=1}^{m} n_k p_k(x)}. \quad (18) \]
The result matches equation (16) with \( n_j = n\alpha_j \) and \( \beta_j = 0 \).

Their cutoff heuristic takes
\[ w_j(x) \propto n_j p_j(x) 1_{n_j p_j(x) \geq \gamma \max_k n_k p_k(x)}, \quad (19) \]
for some \( 0 \leq \gamma \leq 1 \), and their power heuristic takes
\[ w_j(x) \propto (n_j p_j(x))^\rho, \quad (20) \]
for \( \rho \geq 0 \). In both heuristics the weights are normalized to sum to unity. Sending \( \rho \to \infty \) in the power heuristic or taking \( \gamma = 1 \) in the cutoff heuristic gives rise to the maximum heuristic
\[ w_j(x) \propto 1_{n_j p_j(x) = \max_k n_k p_k(x)}. \quad (21) \]
Unless there are ties among the \( n_j p_j \), equation (21) puts all of the weight on one of the \( j \)'s.

6 Examples

This Section presents simulation results for Examples 1 and 2 described in Section 2. Both examples have an importance sampling density that when scaled properly is visually indistinguishable from the integrand. Despite this, Example 1 has infinite variance under importance sampling, while Example 2 has very small variance under importance sampling. Defensive importance sampling cures the problem of Example 1 while losing the accuracy in Example 2.
We compare 11 methods. For each, we compute an estimate of \( I \) using \( n = 10^5 \) observations. Then using 20 independent replicates we compute

\[
Q = \frac{n}{20} \sum_{k=1}^{20} (\hat{I}_k - I)^2,
\]

where \( I \) is the true integral value and \( \hat{I}_k \) is the estimate in the \( k \)'th replicate. For methods with negligible bias and finite variance, (22) estimates the asymptotic variance. Different methods were based on independent samples.

Here are the methods we compared:

**IID**: \( X_1, \ldots, X_n \) are iid from \( q = U(0, 1)^5 \), and \( \hat{I} = \hat{I}_q = n^{-1} \sum_{i=1}^{n} f(X_i) \).

**IS**: \( X_1, \ldots, X_n \) are iid from \( p(x) \), and \( \hat{I} = \hat{I}_p = n^{-1} \sum_{i=1}^{n} f(X_i)/p(X_i) \).

**DIS**: \( X_1, \ldots, X_n \) are iid from \( p_\alpha = \alpha_1 q + \alpha_2 p \), and \( \hat{I} = \hat{I}_{\alpha} \).

**MCV**: Use \( \hat{I}_{\alpha,\beta} \) of (16), with \( p_1 = q, p_2 = p, n_j = n \alpha_j \).

**BAL**: Use \( \hat{I}_{n,w} \) of (17) with \( p_1 = q, p_2 = p, n_j = n \alpha_j \), and \( w \) from (18).

**CUT**: Like BAL, except \( w \) is from (19), with \( \gamma = 0.1 \), and \( n_j = n/2 \).

**POW**: Like BAL, except \( w \) is from (20), with \( \rho = 2 \), and \( n_j = n/2 \).

**MAX**: Like BAL, except \( w \) is from (21), and \( n_j = n/2 \).

Of the 8 methods above, DIS, MCV and BAL are investigated at \( \alpha_1 = 0.1 \) and at \( \alpha_1 = 0.5 \), bringing the total to 11 methods. These methods are closely related: MCV without control variates is BAL, and BAL with random mixing is DIS.

The results for these examples are plotted in Figure 2. A reference rectangle there, based on an approximate \( F_{20,20} \) distribution for \( Q \) ratios, can be used to guage statistical significance.
Figure 2: Shown are normalized mean squared errors $Q$ from Equation (22), for Examples 1 and 2 for the 11 sampling methods described in the text. The horizontal axis is for Example 1, and the vertical axis is for Example 2. Statistical significance may be assessed by the plotted rectangle. Points that differ vertically by more than it’s height differ significantly at about the 1% level for Example 2. It’s width has a similar interpretation for Example 1.

As expected, IID sampling does badly on both of these spiky integrands. Also, as expected, importance sampling does very badly on Example 1, despite having matched the mode well, but does very well on Example 2.

DIS with either value of $\alpha_1$ does very well on Example 1, providing the protection it was designed to. But DIS does badly on Example 2 compared to nondefensive IS.

As predicted by Theorem 2, MCV does nearly as well as the better of IS and DIS on both examples. The choice of $\alpha_1$ is of secondary importance compared to the choice of method. We believe that as long as no $\alpha_j$ is too small, that there is little to gain from optimizing over $\alpha$, that cannot be
gained by estimating $\beta$.

The various heuristics from Veach & Guibas (1995) are roughly equivalent on these examples.

7 Positivisation

Here we show that (multiple) importance sampling can achieve zero variance for integrands of mixed sign. Then we combine this positivisation method with mixture sampling and control variates.

7.1 Simple positivisation

For $f$ having mixed sign, write $f = f_+ - f_-$ where $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$. Then $I = \int f_+(x)q(x)dx - \int f_-(x)q(x)dx$. By taking a sample of size $n_+$ from $p_+ \propto f_+q$ and a sample of size $n_-$ from $p_- \propto f_-q$ it is possible to attain a zero variance estimate:

$$
\hat{I}_\pm = \frac{1}{n_+} \sum_{i=1}^{n_+} \frac{f_+(X_{i,+})q(X_{i,+})}{p_+(X_{i,+})} - \frac{1}{n_-} \sum_{i=1}^{n_-} \frac{f_-(X_{i,-})q(X_{i,-})}{p_-(X_{i,-})}.
$$

(23)

There is a practical difficulty in finding $p_\pm$ to attain or approach the optima, but there is no difficulty in evaluating $f_\pm$ at a given value of $x$.

The decomposition above “splits the integrand at 0”. We could as easily write $f(x) = c + (f(x) - c)_+ - (f(x) - c)_-$ for some well chosen constant $c$. In a personal communication, Bennett Fox described a strategy using a value $c < \inf_x f(x)$ so that $(f(x) - c)_- = 0$. One applies importance sampling to $(f(x) - c)_+ = f(x) - c$ and adds $c$ to the result.

More generally, suppose that $h(x)$ is a function for which $\int h(x)q(x)dx =$
\( \mu \) is known. Let \( X_{i,\pm} \) be independent from \( p_{\pm} \), \( i = 1, \ldots, n_{\pm} \), and

\[
\hat{I}_{h_{\pm}} = \mu + \frac{1}{n_{+}} \sum_{i=1}^{n_{+}} \frac{(f - h)_{+}(X_{i,\pm})}{p_{+}(X_{i,\pm})} - \frac{1}{n_{-}} \sum_{i=1}^{n_{-}} \frac{(f - h)_{-}(X_{i,\pm})}{p_{-}(X_{i,\pm})}.
\] (24)

Here and elsewhere, an expression like \( [(f - h)_{+} q](X) \) replaces \( (f(X) - h(X))_{+} q(X) \) in order to shorten formulas. The estimate (24) is unbiased if \( p_{\pm} > 0 \) when \( (f - h)_{\pm} q > 0 \) respectively. The ideal densities \( p_{\pm} \) are now proportional to \( (f - h)_{\pm} q \).

The function \( h \) is like a control variate, though it has a known nominal expectation, not a known integral, and it is used in a nonstandard way. A good candidate for \( h \) would be a function that was close to \( f \) over most of \( D \) and for which one can guess where the greatest differences are likely to be, in order to target those regions with densities \( p_{\pm} \).

Another way to split the integrand, using \( h \) with known integral \( \int h(x) dx = \mu \), is to write

\[
\tilde{I}_{h_{\pm}} = \mu + \frac{1}{n_{+}} \sum_{i=1}^{n_{+}} \frac{(f q - h)_{+}(X_{i,\pm})}{p_{+}(X_{i,\pm})} - \frac{1}{n_{-}} \sum_{i=1}^{n_{-}} \frac{(f q - h)_{-}(X_{i,\pm})}{p_{-}(X_{i,\pm})}.
\] (25)

For a uniform nominal distribution, with \( q(x) = 1 \), we find \( \tilde{I}_{h_{\pm}} = \hat{I}_{h_{\pm}} \), but in general they are different. Effective use of \( \tilde{I}_{h_{\pm}} \) can be made using a function \( h \) with known integral, for which \( f q - h \) is small in most places, assuming that one can guess where the greatest differences are likely to be.

We prefer (24) to (25) because we think it will be easier to select \( h \) to approximate \( f \) than to approximate \( f q \). For this reason, generalizations of positivisation below are generalizations of (24), not (25).
7.2 Partition of identity

Importance sampling can be applied with quasi-Monte Carlo (Niederreiter 1992) sampling, or other methods that benefit from smooth integrands. But the functions \((f - h)_{\pm}\) are not necessarily smooth even if \(f(x)\) and \(h(x)\) are both smooth. It is thus of interest to positivise \(f - h\) without losing smoothness. (This discussion does not apply to discrete \(X\) where differentiability of \(f\) is not advantageous.)

Define a partition of the identity by a set of functions \(v_j, \ j = 1, \ldots, r\) satisfying

\[
z = \sum_{j=1}^r v_j(z), \quad -\infty < z < \infty. \tag{26}
\]

If also, for each \(j\) we either have \(v_j(z) \geq 0\) for all \(z\), or \(v_j(z) \leq 0\) for all \(z\), we call these functions a semi-definite partition of the identity function. A smooth semi-definite partition of identity can be achieved by

\[
\frac{z}{2} \pm \sqrt{\eta + \left(\frac{z}{2}\right)^2},
\]

where \(\eta > 0\).

Our estimate now becomes

\[
\hat{I}_{h,v} = I_h + \sum_{j=1}^r \frac{1}{n_j} \sum_{i=1}^{n_j} v_j \left( ((f - h)(X_{ji})) q(X_{ji}) \right) \frac{1}{p_j(X_{ji})} \tag{27}
\]

where \(X_{ji}\) are independently sampled from the density \(p_j\). The ideal \(p_j\) are proportional to \(|v_j(f - h)|q\).

7.3 Mixture sampling and positivisation

Mixture sampling can be combined with positivisation by applying mixture sampling with control variates to each of the \(r\) terms in (27). In principle a
different mixture sampling method, with control variates could be applied to each of the \( r \) terms.

A great simplification occurs if we use the same mixture density and control variates for all \( r \) terms, and use a common set of data points \( X_{ji} \) for all \( r \) integrals. The \( v_j \) recombine to form the identity and instead of \( mr \) control variate coefficients, we only need \( m \) of them.

With independent \( X_i \sim p_a \), the resulting estimator is

\[
\tilde{I}_{h,a,\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{[(f - h)q(X_i) - \sum_{k=1}^{m} \beta_k p_k(X_i)]}{p_a(X_i)} + \mu + \sum_{j=1}^{m} \beta_j, \tag{28}
\]

and using deterministic mixtures it is

\[
\hat{I}_{h,a,\beta} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{[(f - h)q(X_{ji}) - \sum_{k=1}^{m} \beta_k p_k(X_{ji})]}{p_a(X_{ji})} + \mu + \sum_{j=1}^{m} \beta_j, \tag{29}
\]

where \( X_{ji} \sim p_j \) are independent, for \( i = 1, \ldots, n_j \), and \( j = 1, \ldots, m \). The coefficients \( \beta_j \) are estimated by regression of \( (f - h)q/p_a \) on predictors \( p_k/p_a \), replacing \( f \) by \( f - h \) in the discussions of Sections 3 and 4.1.

We suggest the following strategy in applications, where subject matter knowledge allows it. First find a suitable proxy function \( h \) that is close to \( f \) and has known integral. Then find densities \( p_1 \) and \( p_2 \) that are large where \((f - h)_+ \) and \((f - h)_- \) are, respectively. Then take a third density \( p_3 \) as the nominal or some other defensive density.

This procedure is illustrated by Example 4 in Section 8. Where \((f - h)_\pm \) has a small number of modes with known locations, a mixture density can be constructed for each mode. If a more general partition of identity having \( r \) terms is used, as in equation (26), one or more component densities can be designed for each such term.

In equations (28) and (29), \( h \) and \( \mu \) appear with a coefficient of 1. We could also estimate a coefficient for them, or indeed, replace \( h \) by a linear
combination of several control variates while replacing $\mu$ by the same linear combination of their integrals.

8 Positivisation example

This example is modeled after some integrands in computational finance, where $\int f(x)q(x)dx$ represents the value of a financial instrument. Suppose that $\int h(x)q(x)dx = \mu$ and that $f$ is equal to $h$ subject to a floor of $A$ and a ceiling of $B$. The value $\int h(x)q(x)dx$ might be known in closed form or be obtainable by some fast algorithm.

Example 4 illustrates the essence of this setting, with functions much simpler than those arising in finance.

**Example 4** Take $\mathcal{D} = (0, 1)^5$, $q = U(0, 1)^5$. Let $h(x) = 100 \left( \sum_{j=1}^{5} x^j - 1 \right)$ and $f(x) = \max(\min(h(x), 300), -25)$. Under $p_1$, the components $X^j$ are independent $N(0, \sigma_1^2)$ variables conditioned to lie in $(0, 1)$. Under $p_2$, the components $X^j$ are independent $N(1, \sigma_2^2)$ variables conditioned to lie in $(0, 1)$. The third density is $p_3 = q = U(0, 1)^5$. We take $\sigma_2 = 0.2236$ and $\sigma_1 = 0.75 * \sigma_2 = 0.1677$.

Trivially $\mu = 150$, and using a result on the volume of a simplex we find $I = 150 - 100/6! + (0.75)^5 * 75/6! = 149.8858$.

We expect that in practice, subject matter knowledge will often allow one to know roughly where the spikes in $f - h$ are, perhaps because $h$ is known to be monotone in its arguments. But we also expect that it will be hard to find a densities that precisely match the spikes. Our densities $p_1$ and $p_2$ are meant to mimic this qualitatively correct, but imperfect, knowledge. Our choices of $\sigma_1$ and $\sigma_2$ give rise to some $X$’s falling outside the spike regions and into the $f - h = 0$ region, while at the same time failing to cover the
<table>
<thead>
<tr>
<th>IID</th>
<th>CCV</th>
<th>PM</th>
<th>PMDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$4.00 \times 10^3$</td>
<td>$6.07 \times 10^0$</td>
<td>$1.90 \times 10^3$</td>
<td>$5.49 \times 10^{-2}$</td>
</tr>
<tr>
<td>MCV-R</td>
<td>MCV-Rh</td>
<td>MCV-D</td>
<td>MCV-Dh</td>
</tr>
<tr>
<td>$4.47 \times 10^{-2}$</td>
<td>$5.22 \times 10^{-2}$</td>
<td>$2.66 \times 10^{-2}$</td>
<td>$8.17 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Table 1: The normalized mean squared errors (22) are given for the simulation of Example 4.

... certain corners of the spike regions. This should cause naive positivisation of $f - h$, without defensive sampling, to fail.

For the positivisation methods, let $p_- = p_1$ and $p_+ = p_2$. For the mixture methods we use $\alpha_1 = \alpha_2 = 0.45$, and $\alpha_3 = 0.1$. This makes relatively light use of the defensive density $p_3 = q$. We considered these methods:

**IID:** $X_1, \ldots, X_n$ are iid from $q = U(0, 1)^5$, and $\hat{I} = \hat{I}_q = n^{-1} \sum_{i=1}^n f(X_i)$.

**CCV:** Classical control variates (10), using $h$ and $X_i \sim q$.

**PM:** The positivisation method (24) with $n_\pm = n/2$.

**PMDM:** The positivisation method (24) replacing $p_\pm$ by defensive mixtures $0.1q + 0.9p_\pm$ and using control variates.

**MCV-R:** Mixture sampling (28) with density control variates.

**MCV-Rh:** MCV-R with an estimated coefficient for $h$ and $\mu$.

**MCV-D:** MCV-R using deterministic mixture sampling as at (29).

**MCV-Dh:** MCV-Rh with deterministic mixture sampling.

For each method, we conduct 20 independent runs, each with sample size $n = 10^5$ and compute $Q$ from (22). The results are reported in Table 1.
Using \( h \) as a control variate reduces the variance by roughly a factor of 650, compared to plain IID sampling. The MCV methods reduced the variance still further by amounts ranging from nearly 75 (MCV-Dh) to nearly 230 (MCV-D). For statististical significance at the 1% level, a variance ratio needs to be larger than 3.32.

The naive PM method failed as expected. The rough qualitative accuracy of \( p_1 \) and \( p_2 \), that was so well exploited by the MCV methods, is not sufficient without some defensive mixing. The PMDM method was much better than the PM method, and was competitive with the MCV methods.

**Acknowledgements**

This work was supported by the National Science Foundation. We thank Bennett Fox, Paul Glasserman, Philip Heidelberger, Tim Hesterberg and Eric Veach for discussions. We also thank all the reviewers for comments that greatly improved the presentation.

**References**


